Learning Sparse and Discriminative Multimodal Feature Codes for Finger Recognition

Shuyi Li, Bob Zhang, *Senior Member, IEEE*, Lunke Fei, *Member, IEEE*, Shuping Zhao, *Member, IEEE*, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*—Compared with uni-modal biometrics systems, multimodal biometrics systems using multiple sources of information for establishing an individual's identity have received considerable attention recently. However, most traditional multimodal biometrics techniques generally extract features from each modality independently, ignoring the implicit associations between different modalities. In addition, most existing work uses hand-crafted descriptors that are difficult to capture the latent semantic structure. This paper proposes to learn the sparse and discriminative multimodal feature codes (SDMFCs) for multimodal finger recognition, which simultaneously takes into account the specific and common information among inter-modality and intra-modality. Specifically, given the multimodal finger images, we first establish the local difference matrix to capture informative texture features in local patches. Then, we aim to jointly learn discriminative and compact binary codes by constraining the observations from multiple modalities. Finally, we develop a novel SDMFC-based multimodal finger recognition framework, which integrates the local histograms of each division block in the learned binary codes together for classification. Experimental results on three commonly used finger databases demonstrate the effectiveness and robustness of the proposed framework in multimodal biometrics tasks.**

***Index Terms*—Finger recognition, sparse and discriminative feature, binary codes, inter-modality, intra-modality.**

I. Introduction

UNI-MODAL biometrics generally identifies an individual based on a single information source from his/her physiological or behavioral characteristics, such as face [1], [2], palmprint [3], gait and voice [4], etc. Unfortunately, these uni-modal biometrics systems are often limited by a variety of problems such as noise in sensed data, spoofing attacks, intra-class variations, and inter-class similarities [5]. Therefore, multimodal biometrics technologies, which can integrate more complementary information presented by multiple sources, have received extensive attention in practical applications [6], [7].

Due to its universality and high accuracy, finger-based traits, such as finger-vein (FV) [8], fingerprint [9], and finger-knuckle-print (FKP) [10], [11], exhibit remarkable advantages in identity authentication. Among them, finger-vein and finger-knuckle-print patterns have plentiful texture features and are located in close proximity of a finger. Over the past decade, various finger-based multimodal recognition technologies have been exhaustively investigated and become increasingly significant because of their convenience [9], [13]. Generally speaking, a completed multimodal biometrics system is composed of region of interest (ROI) extraction, feature representation, feature fusion, and matching. The purpose of ROI extraction is to detect a sub-region and remove the redundant information from the complex background in the initial captured image. Feature representation aims to extract discriminative features to make the data from different classes more separable, and feature fusion is to design effective strategies to integrate the characteristics of multiple modalities. Matching is to classify the extracted features by using an appropriate classifier. In such multimodal biometrics systems, a suitable feature representation approach and fusion strategy is extremely important for improving the recognition performance. According to different fusion levels, the fusion strategies can be roughly separated into pixel-level fusion, feature-level fusion, score-level fusion, and decision-level fusion [12]. It has been proven that the feature-level fusion is capable of achieving more effective performances [14]. For example, Yang *et al.* [15] presented a Weber's law-based cross section asymmetrical coding feature-level fusion algorithm, obtaining a satisfactory recognition performance in bi-modal finger recognition.

Existing finger-based feature representation approaches can be broadly grouped into holistic feature-based methods and local feature-based methods. The holistic feature-based methods, such as principal component analysis (PCA) [16] and linear discriminant analysis (LDA) [17], usually convert the original data into a low-dimensional subspace such that the projection features have more discriminant capability. Representative local feature-based descriptors include local binary pattern (LBP) [18], and Gabor filters [19]. However, most of the local feature-based descriptors are manual designed and generally require much professional knowledge. Moreover, for different modalities and different databases, the hand-designed feature

Manuscript received 23 June 2021; revised 15 September 2021; accepted 18 November 2021. Date of publication 2 December 2021; date of current version 9 March 2023. This work was supported in part by the University of Macau under Grant MYRG2018-00053-FST, in part by the National Natural Science Foundation of China under Grant 61602540, and in part by the National Natural Science Foundation of China under Grant 62176066. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jialie (Jerry) Shen. *(Corresponding author: Bob Zhang.)*

Shuyi Li and Bob Zhang are with the PAMI Research Group, Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yb97443@um.edu.mo; bobzhang@um.edu.mo).

Lunke Fei and Shuping Zhao are with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: flksxm@126.com; yb77458@um.edu.mo).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@um.edu.mo).

Digital Object Identifier 10.1109/TMM.2021.3132166

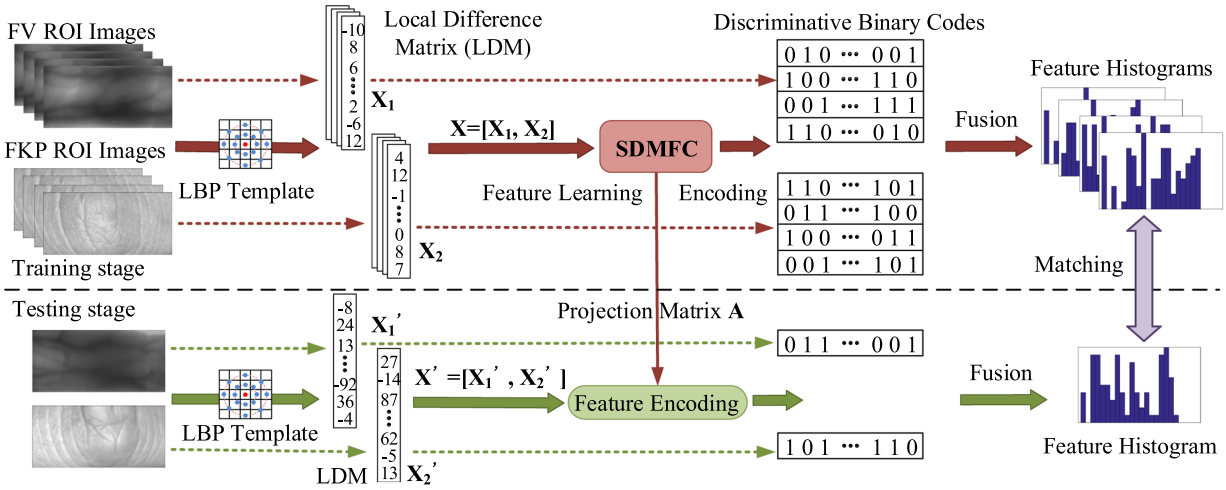


Fig. 1. The basic idea of the proposed SDMFC framework for multimodal finger recognition. 1) For the training FV and FKP images, we first apply the LBP template to extract the local difference matrices (LDMs). 2) Following this, SDMFC aims to jointly learn a set of projection functions that map and encode the LDMs into discriminative binary codes. 3) Given the testing images, we also establish the LDMs and encode them to binary features via the pre-trained projection matrix. 4) Lastly, we concatenate the feature histograms of each modality for matching.

descriptors may not be effective to extract the dominant features. To this end, the convolutional neural networks (CNN)-based methods have attracted much interest in finger biometric recognition [20], [21]. Due to the fact that existing finger biometric datasets have small sample sizes, the CNN-based methods may be limited in practical applications. Therefore, most of the CNN-based approaches usually explore to fine-tune the network parameters by using a pre-trained model (such as VGGNet [22] and ResNet [23]) on their considered datasets [24]. Beyond that, the feature learning-based methods, including subspace learning [25], [26], dictionary learning [27], [28], sparse representation [29]–[31], hashing learning [32], [33], and binary code learning [3], [11], [34] have been proposed in recent years. Among them, the binary code learning-based methods produce excellent performances in the field of finger and other biometrics identification. For example, Liu *et al.* [35] proposed a cross-modality binary code learning algorithm to map features from multiple modalities into a new subspace with compact binary codes. Fei *et al.* [36] developed a multi-view feature learning method to project the texture and direction features into binary codes, which achieved promising results in finger-knuckle-print and palmprint recognition.

In order to perform multimodal finger recognition effectively, it is critical and fundamental to solve the core research problems: 1) How to optimally represent the correlation between different modalities? 2) How to model a data-independent feature learning framework to adaptively learn discriminant multimodal features? With the demand for higher accuracy and the emergence of complex data, these problems are becoming increasingly important and have not been solved properly in previous studies. In this paper, we study these problems systematically and comprehensively, by proposing a sparse and discriminative multimodal feature codes (SDMFCs) learning method to guide the feature learning procedure adaptively. The proposed SDMFC can be directly applied to various authentication applications, such as bank ATM, access control, PC login, electronic payment, and so on. To the best of our knowledge, this work is indeed groundbreaking in trying to solve the above two problems in multimodal finger recognition. We believe that this work will have a significant impact on the literature related to multimodal biometrics, especially given that the real applications of authentication in multimedia computing are booming.

Fig. 1 illustrates the flowchart of the proposed SDMFC framework. Given the training finger-vein and finger-knuckle-print samples, we first established the local difference matrices (LDMs) of each modality by using the LBP template. Afterwards, we integrated the LDM features of each modality as the input of SDMFC for training. Next, the proposed SDMFC jointly learned a set of linear mapping functions that project and encode the raw multimodal features into binary codes such that the projected features are more compact and discriminative. At last, we calculated the feature histograms of each modality and concatenated them together to form the final feature representation. In the training stage, the projection functions are learned and saved in advance, and can be directly used to encode the test samples in the testing stage. For the given test samples, we first calculated the LDMs and then mapped them into compact binary codes by the pre-learned projection matrix. Following this, we computed the block-wise histograms of these test samples and integrated them for matching.

Overall, the main contributions of our work are highlighted as follows:

- A novel sparse and discriminative multimodal feature learning method with shared structural space is proposed to jointly learn compact binary codes for multimodal finger recognition. The latent correlation between the multimodal features is captured by narrowing the distance of the inter-modality samples with the same semantic label.
- The proposed SDMFC aims to transform finger features from multiple modalities to a common space, and perform efficient feature fusion in the projection space to exploit the common and specific information between

the inter-modality and intra-modality samples. Experimental results show that integrating finger-vein and finger-knuckle-print features can significantly boost the recognition performance.

- We conducted extensive experiments on three widely used multimodal finger datasets where the results demonstrate the effectiveness of the proposed method in terms of both accuracy and efficiency. Without loss of generality, the proposed SDMFC can be easily extended and applied in other multi-biometrics tasks, such as face and fingerprint fusion, palmprint and palm vein fusion, to name a few.

The paper is organized as follows: In Section II, we briefly review the binary features-based methods and multimodal recognition methods. Section III presents the proposed SDMFC and its optimization process in detail. The obtained experimental results are discussed and analyzed in Section IV. Finally, the conclusions are drawn in Section V.

II. Related Work

A. Binary Features Representation

Due to its great robustness to local changes, such as illumination changes and rotation variations, the binary features-based representation methods have gained tremendous research momentum. Representative binary features representation methods including LBP [18] and its variant [37], as well as local graph structure-based (LGS) [38], have produced outstanding results in finger biometric recognition. While binary features such as LBP and LGS-like features have been used in finger-based biometric recognition, most of them are hand-crafted and require strong prior knowledge. In recent years, works that utilize binary code learning for feature representation have gradually become a research hotspot. For example, Lu *et al.* [39] introduced a compact binary face descriptor to automatically learn binary features for facial recognition. Afterwards, in [40], a local binary feature learning method was explored to jointly learn a set of projection functions to transform the facial data into discriminative binary codes. Fei *et al.* [41] developed a discriminant direction binary palmprint descriptor (DDBPD) that converts the palmprint direction features into binary features. These show that the binary code learning-based approaches have obtained attractive recognition results in biometrics recognition tasks. In this work, we propose to adaptively learn a common latent space and project the multimodal features into discriminative binary codes.

B. Multimodal Analysis

Multimodal data from multiple sources, such as images, text, and video, are semantically correlated and provide complementary information to each other [42], [43]. With the rapid development of multimedia information, such multimodal systems combining heterogeneous data from various sensors have been extensively explored. For example, Zhang *et al.* [44] integrated the inter-correlations between images and text information for marketing intent analysis. Elmadany *et al.* [7] proposed to learn a discriminative common space between two different modalities from RGB videos for human action recognition. Wang *et al.* [33] introduced a multimodal hashing method to transform heterogeneous data into latent semantic spaces for cross-modal similarity search tasks. Although multimodal methods have achieved impressive performances in various scenarios, joint multimodal learning has been rarely addressed in multimodal finger recognition, with a few exceptions [12], [13]. Li *et al.* [12] proposed a joint feature learning (JDFL) method to describe the correlations among multiple modalities for multimodal finger recognition. Furthermore, a sparse coding based feature learning algorithm, called JDSC, was presented in [13] and applied for hand-based multimodal recognition.

Recently, a variety of multimodal finger recognition technologies have been proposed and widely used in identity authentication scenarios. For example, Yang *et al.* [45] proposed a comparative competitive coding-based (Compcode) fusion method that integrated the finger-vein and finger-dorsal-texture features. Yang *et al.* [46] developed a cancelable multi-biometrics system by extracting the minutia-based fingerprint features and the image-based finger-vein features, respectively. In addition, Zhang *et al.* [47] established a graph structure based feature-level fusion model (Graph_fusion) to characterize the tri-modal finger images. A generalized symmetric LGS (GSLGS) descriptor was designed in [38] to independently explore the tri-modal finger features and perform fusion by concatenating the histograms of each modality. Previous studies in the field of multimodal finger recognition are plagued by problems that have not been addressed properly. 1) Most of the existing methods usually focus on extracting information from multiple modalities separately, while the latent common information among the inter-modality samples from the same class is usually ignored. 2) Most of the conventional multimodal recognition methods are hand-designed. However, relying solely on the features extracted by hand-craft may be ineffective at representing the dominating discrimination of multiple modalities. Hence, how to optimally express the common and specific information between multiple modalities remains a main challenge in multimodal biometrics tasks. To address these problems, we present a sparse and discriminative multimodal feature codes (SDMFCs) learning method, which considers the relationship of the intra-class and inter-class data, as well as the association of the intra-modality and inter-modality data.

Note that although DDBPD, JDFL, JDSC, and SDMFC all use feature learning in its learning procedure, their main concepts are different. First, unlike DDBPD that learns features from a single modality, SDMFC jointly utilizes the intra-modality and inter-modality data from multiple modalities along with their label information. Secondly, different from JDFL and JDSC that perform feature learning from the direction information, SDMFC learns multimodal features from the informative local texture features. Secondly, SDMFC jointly learns a common projection matrix for all modalities, while JDSC learns different projection matrices for different modalities. In addition to the intra-modality and inter-modality constraints in JDFL, SDMFC also combines a projection error constraint and a sparse norm constraint to guarantee the learned binary features is more discriminative and sparse. Therefore, the proposed SDMFC is different from other multimodal feature learning methods.

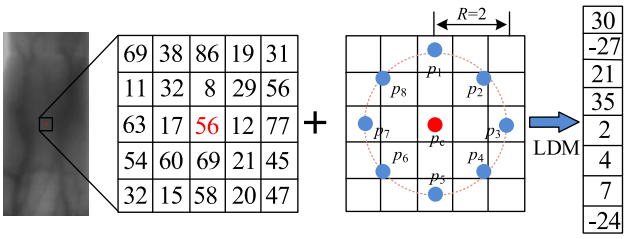


Fig. 2. An illustration to show how to extract the local difference matrix.

III. SDMFC-Based Multimodal Finger Recognition

In this section, we present the details of the proposed SDMFC. First, a local difference matrix (LDM) is formed to describe the local texture features of finger images. Then, the overall objective function of SDMFC and the theoretical analysis are given. Finally, a fusion strategy based on SDMFC is developed for multimodal feature description.

A. Local Difference Matrix

As is well known, both finger-vein and finger-knuckle-print patterns contain plentiful and distinctive texture features. As one of the most powerful texture feature descriptors, LBP captures the signs among the center pixel and the neighbors. Based on the LBP template, we form a local difference matrix to describe the local texture features of the finger images. Specifically, for each center pixel of a finger-vein (or finger-knuckle-print) image, we selected the neighboring pixels in the local patch with the size of $(2R{+}1){\times}(2R{+}1)$, where R is the neighborhood radius. Afterwards, we calculate the difference responses between the central pixel and its neighborhood pixels in sequence. Fig. 2 describes how to extract the LDM from a finger-vein image. As shown in Fig. 2, R is selected as 2 and we selected eight neighboring pixels through experiments to form LDM, such that LDM is composed of an 8-dimensional feature vector for each pixel.

B. Learning of SDMFC

Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m]$ is a set of multimodal training data, where $\mathbf{X}_m = [x_{m,1}, x_{m,2}, \ldots, x_{m,N}] \in R^{k\times N}(1 \le m \le M)$ be the data matrix of the m-th modality and $x_{m,n}(1 \le n \le N)$ is a LDM extracted from a finger image. M is the total number of modalities and N denotes the total number of training samples. k represents the dimension of LDM. Table I summarizes the notations used and its corresponding descriptions.

In this subsection, we design a linear sparse and discriminative feature learning method, aiming to project the LDMs of multimodal finger images into a common subspace. Suppose that SDMFC learns k mapping functions, which project each $x_{m,n}$ into a binary matrix $y_{m,n} = [y_{m,n,1}, \ldots, y_{m,n,k}]$ with $y_{m,n,k} \in \{0,1\}^{k\times 1}$. Then, the k-th mapping function is defined as follows:

$$y_{m,n,k} = \mathrm{sgn}(a_k^T x_{m,n}) \tag{1}$$

where $y_{m,n,k}$ represents the learned k-th binary code, $\mathrm{sgn}(\cdot)$ is the element-wise sign function, a_k denotes the learned projection vector of the k-th mapping function.

TABLE I
The Notations Used and Its Corresponding Descriptions

Notations	Descriptions
$x_{m,n}$	LDM of n-th image on the m-th modality
X_m	training data of the m-th modality
X	combined data matrix of multiple modalities
a_k	the learned k-th projection function
A	the projection matrix
$y_{m,n,k}$	the projected k-th features of $x_{m,n}$
$y_{m,n}$	binary matrix of multiple $y_{m,n,k}$
Y_m	projected binary matrix of the m-th modality
Y	combined binary matrix of multiple modalities
R	the neighborhood pixels forming LDM
M	total number of modalities
N	total number of training samples
k	length of binary codes
λ_1, λ_2, λ_3	balance parameters for each constraint
$\mathbf{\Gamma}$	the sub-dataset of within-class samples
$\mathbf{\Lambda}$	the sub-dataset of between-class samples
S	label matrix of sample relationships
I	maximum iteration number
σ	convergence parameter

For the projected features, we minimize the quantization error to preserve the raw semantic information. Then, we enforce the sparse norm constraint on the projection matrix, such as l_1-norm and $l_{2,1}$-norm, to make the mapped features more sparse. For the intra-modality data, the distance of the within-class samples is minimized, at the same time, the distance of the between-class samples is maximized. Lastly, since the inter-modality samples from the same class have the same semantic label, the distance between them is minimized in the common projection space. Therefore, the overall objective function can be formulated as:

$$\begin{aligned}\min_{a_k} &\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}\|y_{m,n,k} - a_k^T x_{m,n}\|^2 + \lambda_1\|a_k\|_{2,1} \\ &+ \lambda_2\Big(\sum_{i,j\in\mathbf{\Gamma}}\|y_{m,i,k} - y_{m,j,k}\|^2 - \sum_{i,j\in\mathbf{\Lambda}}\|y_{m,i,k} - y_{m,j,k}\|^2\Big) \\ &+ \lambda_3\|y_{m,n,k} - y_{m+1,n,k}\|^2, \text{subject to } a_k a_k^T = 1^{k\times 1},\end{aligned} \tag{2}$$

where $\mathbf{\Gamma}$ denotes the sub-dataset containing the within-class images, and $\mathbf{\Lambda}$ is the sub-dataset that consists of the between-class images. If i and j belong to the same class, they are assigned to $\mathbf{\Gamma}$, otherwise, they are assigned to $\mathbf{\Lambda}$. Due to the fact that the $l_{2,1}$ norm has a better row-sparsity and efficient feature selection property than the l_1 norm, the $l_{2,1}$ norm is used here, which can make the learned projection matrix have better interpretability [17].

Let $\mathrm{A} = [a_1, a_2, \ldots, a_k]$ be the projection matrix, then, the LDM features of multiple modalities can be projected into Y as follows:

$$\mathrm{Y} = \mathrm{sgn}(\mathrm{A}^T\mathrm{X}) \tag{3}$$

where Y represents the learned binary features of all training images. Following this, the objective function of (2) can be rewritten as follows:

$$\begin{aligned}\min_{\mathrm{A}} & \sum_{m=1}^{M} \|\mathrm{Y}-\mathrm{A}^T\mathrm{X}\|_F^2+\lambda_1\|\mathrm{A}\|_{2,1} \\ & +\lambda_2\Big(\sum_{i,j\in\mathbf{\Gamma}}\|\mathrm{Y}_{m,i}-\mathrm{Y}_{m,j}\|_F^2-\sum_{i,j\in\mathbf{\Lambda}}\|\mathrm{Y}_{m,i}-\mathrm{Y}_{m,j}\|_F^2\Big) \\ & +\lambda_3\sum_{n=1}^{N}\|\mathrm{Y}_{m,n}-\mathrm{Y}_{m+1,n}\|_F^2, \text{subject to } AA^T=I,\end{aligned} \tag{4}$$

where $\mathrm{X}=[\mathrm{X}_1,\mathrm{X}_2]$, and $\mathrm{Y}=[\mathrm{Y}_1,\mathrm{Y}_2]$. $\mathrm{Y}_{m,n}$ is the projected features of the n-th sample from the m-th modality, $\|\cdot\|_F$ denotes the l_F-norm ('Frobenius' norm), λ_1, λ_2, and λ_3 are three non-negative balance parameters of the corresponding constraints.

C. SDMFC Optimization

To simplify the presentation, in this paper, we first explore the application of SDMFC to bi-modal biometric data. Hence, M is set to 2 in (4). Without loss of generality, our proposed SDMFC can be easily extended to cases with more biometric modalities with rich texture features. The overall objective function of SDMFC combined with these feature learning terms is defined as

$$\min_{\mathrm{A}}\mathcal{F}(\mathrm{A})=\min_{\mathrm{A}}\mathcal{F}_1+\lambda_1\mathcal{F}_2+\lambda_2\mathcal{F}_3+\lambda_3\mathcal{F}_4 \tag{5}$$

where $\mathcal{F}_1$ is the quantization error constraint, $\mathcal{F}_2$ denotes the sparse norm constraint, $\mathcal{F}_3$ is the intra-modality discriminative constraint, and $\mathcal{F}_4$ represents the correlation constraint of the inter-modality samples from the same class.

Specifically, the first constraint of (5) can be solved as follows:

$$\begin{aligned}\mathcal{F}_1&=\|\mathrm{Y}-\mathrm{A}^T\mathrm{X}\|_F^2 \\ &=Tr((\mathrm{Y}-\mathrm{A}^T\mathrm{X})(\mathrm{Y}-\mathrm{A}^T\mathrm{X})^T) \\ &=Tr(\mathrm{A}^T\mathrm{X}\mathrm{X}^TA-2\mathrm{Y}\mathrm{X}^T\mathrm{A}).\end{aligned} \tag{6}$$

The third constraint can be calculated in a matrix form as follows:

$$\begin{aligned}\mathcal{F}_3&=\sum_{m=1}^{2}\sum_{i,j\in\mathbf{\Gamma},\mathbf{\Lambda}}\|(\mathrm{A}^T\mathrm{X}_{m,i}-\mathrm{A}^T\mathrm{X}_{m,j})\mathrm{S}\|_F^2 \\ &=Tr(\mathrm{A}^T\mathrm{X}\mathrm{S}\mathrm{X}^T\mathrm{A})\end{aligned} \tag{7}$$

where $\mathrm{S}\in R^{N\times N}$ denotes a label matrix indicating the category relationships between the within-class and between-class samples. If two images belong to the same class, S is equal to 1, otherwise, S is equal to -1.

The fourth constraint can be written as:

$$\begin{aligned}\mathcal{F}_4&=\|\mathrm{A}^T\mathrm{X}_1-\mathrm{A}^T\mathrm{X}_2\|_F^2 \\ &=Tr(\mathrm{A}^T\mathrm{X}_1{\mathrm{X}_1}^T\mathrm{A}-2\mathrm{A}^T\mathrm{X}_1{\mathrm{X}_2}^T\mathrm{A}+\mathrm{A}^T\mathrm{X}_2{\mathrm{X}_2}^T\mathrm{A}) \\ &=Tr(\mathrm{A}^T(\mathrm{X}_1{\mathrm{X}_1}^T-2\mathrm{X}_1{\mathrm{X}_2}^T+\mathrm{X}_2{\mathrm{X}_2}^T)\mathrm{A}).\end{aligned} \tag{8}$$

Based on the above derivation of (6), (7), and (8), we can obtain the following:

$$\begin{aligned}\mathcal{F}(\mathrm{A})&=Tr(\mathrm{A}^T\mathrm{X}\mathrm{X}^T\mathrm{A}-2\mathrm{Y}\mathrm{X}^T\mathrm{A})+\lambda_1\|\mathrm{A}\|_{2,1} \\ &\quad+\lambda_2Tr(\mathrm{A}^T\mathrm{X}\mathrm{S}\mathrm{X}^T\mathrm{A})+\lambda_3Tr(\mathrm{A}^T(\mathrm{X}\mathrm{X}^T-2\mathrm{X}_1{\mathrm{X}_2}^T)\mathrm{A}) \\ &=Tr(\mathrm{A}^T\mathrm{Q}\mathrm{A})-2Tr(\mathrm{Y}\mathrm{X}^T\mathrm{A})+\lambda_1\|\mathrm{A}\|_{2,1}\end{aligned} \tag{9}$$

with

$$\mathrm{Q}=\mathrm{X}\mathrm{X}^T+\lambda_2\mathrm{X}\mathrm{S}\mathrm{X}^T+\lambda_3(\mathrm{X}\mathrm{X}^T-2\mathrm{X}_1{\mathrm{X}_2}^T) \tag{10}$$

At last, the projection matrix A can be alternately solved by minimizing the following problem:

$$\mathrm{A}^*=\underset{\mathrm{A}}{argmin}\,Tr(\mathrm{A}^T\mathrm{Q}\mathrm{A})-2Tr(\mathrm{Y}\mathrm{X}^T\mathrm{A})+\lambda_1\|\mathrm{A}\|_{2,1} \tag{11}$$

Let the derivative of $\mathcal{F}(\mathrm{A})$ with respect to A be 0, then we obtain

$$\frac{\partial\mathcal{F}(\mathrm{A})}{\partial\mathrm{A}}=2\mathrm{Q}\mathrm{A}-2\mathrm{X}\mathrm{Y}^{\mathrm{T}}+\lambda_1\mathrm{U}\mathrm{A} \tag{12}$$

where U is defined as $\mathrm{U}=\begin{bmatrix}\frac{1}{\|a_1\|_2} & \cdots & 0\\ 0 & \cdots & 0\\ 0 & 0 & \frac{1}{\|a_k\|_2}\end{bmatrix}$, a_k is the k-th row of A, and $\mathrm{A}=\begin{bmatrix}a_1\\ \cdots\\ a_k\end{bmatrix}$. Here, A can be calculated by the following

$$\mathrm{A}=(\mathrm{Q}+\frac{\lambda_1}{2}\mathrm{U})^{-1}\mathrm{X}\mathrm{Y}^{\mathrm{T}}. \tag{13}$$

D. SDMFC-Based Recognition

After the mapping functions are learned, the LDM of each modality can be transformed into k-bit binary codes. To integrate the region-specific of the learned multimodal binary features, we develop a SDMFC-based histogram representation approach for recognition. The detailed procedure of the SDMFC-based recognition is shown in Fig. 3. Specifically, for the learned binary codes, we first calculate the real value of each pixel and obtain the feature map of each modality by weighting and summing the binary codes (see Fig. 3(a) and (b)). Afterwards, the learned feature maps are uniformly divided into non-overlapping divisions with 16×16 size. Section IV-E analyzes the selection of sub-block size. For the sub-blocks in the same position, we integrate the histograms together to form the local feature vectors (see Fig. 3(c)). At last, the final feature histogram is generated by concatenating the local histograms of each sub-block for matching (see Fig. 3(d)). In the matching phase, the feature histograms of two samples are matched by calculating the intersection coefficient [48] to determine the similarity. The proposed SDMFC-based recognition framework is summarized in Algorithm 1.

Algorithm 1: Framework of SDMFC

Training stage

Input: training samples $X=[X_1, X_2]$, parameters λ_1, λ_2, λ_3, iteration number I, convergence parameter σ.

Output: linear projection matrix A.

1: Forming the LDMs of X.
2: Initialize A with a random matrix.
3: Learning the projection matrix A based on the LDMs.
for i=1 to I
Update A by using Eq. (13);
if $\|A^{(i)} - A^{(i-1)}\|^2 < \sigma$ break;
end

Testing stage

Input: testing samples $X'=[X'_1, X'_2]$, learned projection matrix A.

Output: the predicted class of the test image.

4: Forming the LDMs of X'.
5: Calculate binary code by $Y'=\text{sgn}(A^T X')$.
6: Obtain the feature map and its feature histogram.
7: Predict the class of test image.

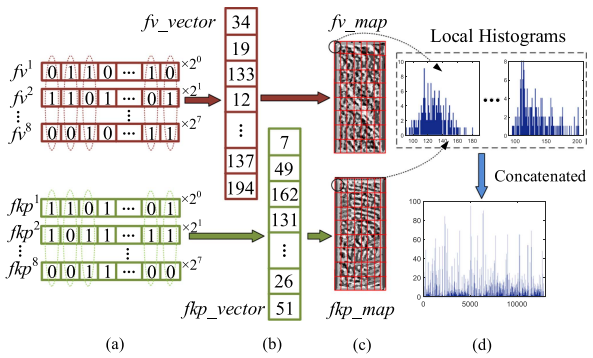


Fig. 3. The detailed procedure of the SDMFC-based recognition. (a) The learned binary codes of the two modalities. (b) These binary codes from each modality are combined into a real value. (c) The feature map of each modality is divided into equally sized sub-blocks. (d) The local feature histograms of each division are calculated and concatenated to establish the final histogram.

IV. Experiments

To evaluate the performance of our proposed SDMFC, we conducted multimodal identification and verification experiments with several state-of-the-art multimodal finger recognition approaches on three multimodal finger databases.

A. Databases

In this section, we used an existing multimodal finger database, Data-multi [12], and set up two multimodal finger databases, SD-PolyU and USM-PolyU.

Data-multi database is a popular multimodal finger database, which captured finger-vein (FV) and finger-knuckle-print (FKP) patterns of a subject, simultaneously. The Data-multi database consists of a FV database (Data-fv) and a FKP database (Data-fkp), which in total contains 11,700 images collected from 585 fingers. Specifically, each finger respectively provided 10 FV images and 10 FKP images. All of the finger ROI images are resized into 90 × 200 pixels in this database.

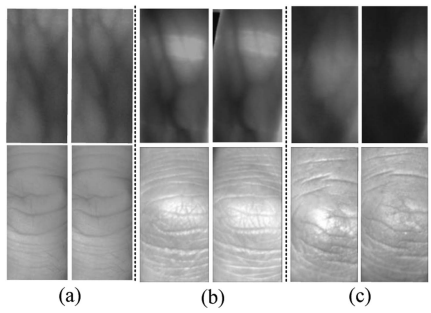


Fig. 4. Some examples of finger ROI images from a subject on the three multimodal finger databases. (a) Data-multi, (b) SD-PolyU, (c) USM-PolyU.

SD-PolyU database is formed through two commonly used uni-modal finger databases, SDUMLA-fv [49] and PolyU-fkp [10]. SDUMLA-fv database contains a total of 636 classes captured from 106 volunteers. Each volunteer provided six fingers of both hands, and each finger was collected six times. Therefore, there are totally 3,816 finger-vein images in the SDUMLA-fv database. For the captured original finger-vein images, we extract the ROI using the inter-phalangeal joint prior method [8], where the size of the ROI image is 110 × 220 pixels. PolyU-fkp database was formed using 660 classes from 165 individuals collected by two separate sessions. Each individual collected four fingers, including the index and middle fingers of the right and left hands, and each finger provided 12 finger-knuckle-print images. To ensure that the categories and sample sizes in this multimodal database are consistent, we selected the first 500 classes and 6 images of each class to setup the used SD-PolyU database.

USM-PolyU database consists of two publicly available unimodal finger databases, FV-USM [50] and PolyU-fkp [10]. FV-USM database involves 492 classes from 123 volunteers comprising of 83 males and 40 females. Every subject provided four fingers and each finger was captured six times in one session. Each individual participated in two sessions, separated by more than two weeks. For these two unimodal finger databases, the first 400 subjects and twelve samples each subject were selected to setup the USM-PolyU database.

Fig. 4 shows some ROI images selected from the three multimodal finger databases. From Fig. 4, we can clearly see that these finger images have obvious intra-class variations in illumination, rotation, and translation, especially in the SD-PolyU and USM-PolyU databases. In order to perform the following feature learning, the ROI of the finger images need to be cropped. In this paper, we uniformly resized the finger images into 45×100 pixels in the Data-multi database, 55×110 pixels in the SD-PolyU database, and 50×150 pixels in the USM-PolyU database, respectively.

B. Evaluation Metrics and Baselines

Two common identification performance metrics, average accuracy (AVE) and standard deviation (STD) are adopted here.

Given a list of accuracy results, its AVE and STD are defined as follows:

$$AVE = \frac{1}{T}\sum_{t=1}^{T}(ACC)_t \tag{14}$$

$$STD = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_i - \overline{x})^2} \tag{15}$$

with

$$ACC = \frac{number\ of\ correctly\ classified\ samples}{total\ number\ of\ samples} \tag{16}$$

where x_i denotes the i-th test accuracy and $\overline{x}$ is the average test accuracy. T represents the total number of tests, in this paper, T is set to 10. Clearly, the larger the AVE is, the greater the recognition result. Also, the smaller the STD is, the better the stability. Furthermore, we also employed two other types of verification performance metrics including equal error rate (EER), and the running time of feature extraction indicating the computational efficiency. Generally speaking, a smaller EER and time cost represent a better verification performance.

To evaluate the performance of the proposed SDMFC, we compared it with several competitors: binary features-based methods (LBP [18], Compcode [45], and GSLGS [38]), sparse representation-based method (E-SRC [30]), deep learning-based methods (VGGNet [22] and ResNet [23]), multimodal finger recognition methods (Compcode [45] and Graph_fusion [47]), as well as the state-of-the-art methods (and DDBPD [41], JDFL [12], and JDSC [13]). LBP is a typical binary features descriptor to express the texture features of an image. E-SRC and DDBPD are two recent feature learning methods for palmprint recognition. VGGNet and ResNet are two popular deep learning models that are widely used in image classification and biometric recognition tasks. In this paper, we utilized these methods to extract the multimodal finger features, separately, before concatenating them at the feature-level for multimodal recognition. Compcode, GLGS, Graph_fusion, JDFL, and JDSC were originally proposed for multimodal finger recognition. For the comparison methods, we implemented them and selected the optimal parameters to obtain the best results. For the proposed SDMFC, we empirically set $\lambda_1 = 1$, $\lambda_2 = 10$, and $\lambda_3 = 0.01$. We present an empirical analysis of the parameter sensitivity, which verifies that SDMFC can achieve the best performance under a suitable parameter setting. Section IV-E details the analysis of the parameters setting for the proposed SDMFC.

C. Results and Analysis

1) Finger Identification: In the following experiments, we randomly selected one sample from each modality of each class for training, and used the remaining samples for testing. Table II depicts the identification results (average accuracy $\pm$ standard deviation) of all methods on the three multimodal finger databases. According to the experimental results, we can observe that the proposed SDMFC obtains a much better performance even in case that a training sample is used. Specifically, the average accuracy of SDMFC achieves 99.8763% on the Data-multi database, 97.4240% on the SD-PolyU database, and 97.7205% on the USM-PolyU database. It is worth noting that the proposed SDMFC outperforms most comparison methods, while obtaining a comparative performance with the state-of-the-art method of JDSC. This result is mainly due to the fact that SDMFC can automatically learn more discriminative multimodal features than the hand-crafted methods as well as fully utilize the specific and common information among multiple modalities. In addition, we tested the effect of different neighborhood radius sizes (R = 2, 3, or 4) for identification. From Table II, we can clearly see that the identification accuracy of SDMFC can be further improved when the neighboring radius are set as 4. Furthermore, the accuracy improvement of the proposed SDMFC on the SD-PolyU and USM-PolyU databases is much greater than that of the Data-multi database. The main reason is that these finger images in the Data-multi database are well-aligned with small illumination changes and rotation variations. This further implies that SDMFC is effective and robust on these databases with obvious illumination and rotation changes.

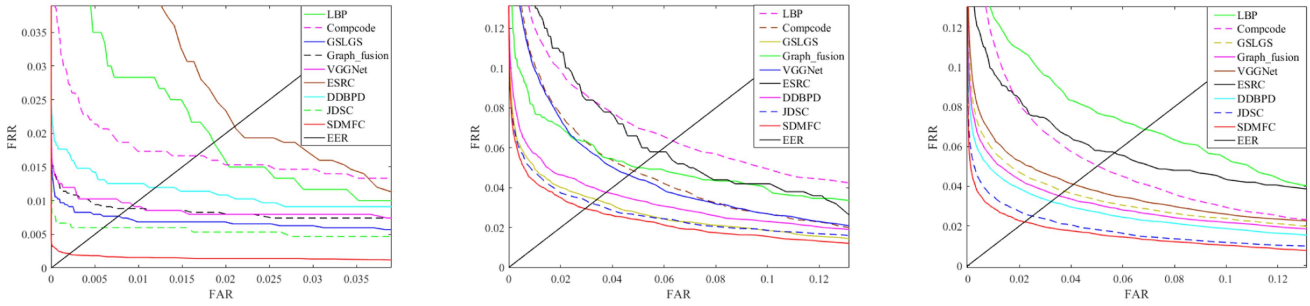
TABLE II
THE AVE AND STD RESULTS (%) OF DIFFERENT METHODS ON THE THREE MULTIMODAL DATABASES

Methods	Data-multi	SD-PolyU	USM-PolyU
LBP	94.3216±0.6247	81.8216±2.0137	76.6741±1.2406
Compcode	96.2985±1.3512	85.2753±2.6427	82.6408±2.1974
GSLGS	98.4784±0.1695	93.0314±0.3635	93.3523±1.0342
Graph_fusion	97.4206±0.6924	89.6227±2.6319	83.7548±2.7318
VGGNet	95.0812±0.6462	90.2683±2.0741	84.1614±2.5102
ResNet	96.3425±0.8126	82.8437±1.0537	88.2219±0.8306
E-SRC	95.7136±0.4853	86.6012±1.2541	80.2147±0.7826
DDBPD	98.6894±0.3617	92.6528±0.6649	93.6682±0.4937
JDFL	99.2982±0.2865	93.2982±0.7895	93.8154±0.6124
JDSC	99.5419±0.3562	**98.7922±0.2439**	97.1738±0.3906
SDMFC(R=2)	99.7930±0.0673	93.7960±2.5288	94.9523±0.2939
SDMFC(R=3)	99.8044±0.1178	94.2620±1.9906	95.4068±0.3850
SDMFC(R=4)	**99.8763±0.0565**	97.4240±1.7629	**97.7205±0.5469**

2) Finger Verification: Furthermore, we followed the aforementioned methods and conducted multimodal finger verification experiments to evaluate the proposed SDMFC. In this subsection, we compared each sample with all other samples and calculated the FAR and FRR of each pair of samples based on the similarity of two matched samples. Fig. 5 shows the receiver operating characteristic (ROC) curves of all methods on the three multimodal finger databases, respectively. From these curves one can find that our SDMFC consistently achieves the lowest EER than the LBP-like descriptor and the state-of-the-art multimodal finger recognition methods. Although the identification performance of the proposed method is slightly inferior to JDSC, its verification performance outperforms JDSC. This is mainly due to the power of the sparse constraint, which makes the learned projection features more sparse and discriminative. This further shows that the proposed SDMFC can fully represent the available information to improve the recognition performance and has the merits of extensive generalization.

D. Computational Efficiency Comparison

The proposed SDMFC optimization problem can be solved by iteratively updating the variable A, where the raw multimodal



(a) (b) (c)

Fig. 5. The ROC curves of different methods on the three multimodal finger databases. (a) Data-multi. (b) SD-PolyU. (c) USM-PolyU.

TABLE III
THE FEATURE EXTRACTION TIME (S) OF DIFFERENT METHODS ON THE THREE MULTIMODAL DATABASES

Methods	Feature extraction time		
	Data-multi	SD-PolyU	USM-PolyU
LBP	0.039	0.041	0.047
Compcode	0.069	0.075	0.086
GSLGS	0.135	0.162	0.169
Graph_fusion	0.564	0.570	0.582
VGGNet	1.430	1.552	1.578
JDFL	**0.032**	**0.034**	**0.036**
SDMFC	0.045	0.052	0.058

TABLE IV
THE AVERAGE IDENTIFICATION ACCURACY (%) VERSUS PARAMETERS λ_1 WHEN λ_2 AND λ_3 ARE FIXED

λ_1	Data-multi	SD-PolyU	USM-PolyU
0.0001	99.7248±0.0352	90.2640±2.4170	95.0240±0.4812
0.001	99.7755±0.0437	90.9840±2.8084	95.0632±0.5064
0.01	99.8367±0.0615	91.1524±1.7806	95.1159±0.2919
0.1	99.8025±0.0240	93.0800±1.8806	96.6591±0.1984
1	**99.8763±0.0565**	**97.4240±1.7629**	**97.7205±0.5469**
10	99.8227±0.0359	95.6280±2.8048	94.8205±0.5355
100	99.7816±0.0615	97.1900±2.1780	96.5136±0.6008
1000	99.8177±0.0836	92.0320±2.2565	95.5955±0.2292
1000	99.8348±0.0714	91.8256±1.8649	95.2306±0.4260

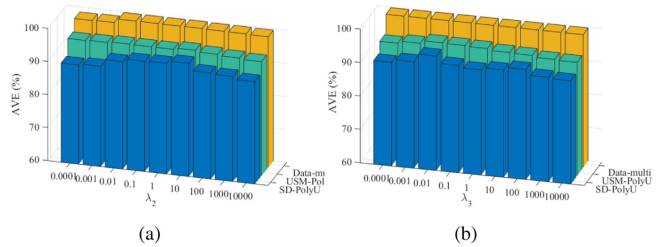
features are encoded into binary features by the pre-learned projection matrix. The coding process only involves matrix calculation and is very fast. Therefore, the computational complexity of SDMFC mainly depends on the learning of the projection matrix A. Given the iteration number I, the computational complexity of updating A is $O(Ik^3)$.

To further test the computational efficiency of our proposed method, we compared the feature extraction time on the three multimodal finger databases. Table III shows the running time of one sample for feature extraction. From Table III, we can find that VGGNet requires the most time for feature learning. More remarkable, the binary features-based approaches, including LBP and JDFL, have higher computational efficiency than the other compared methods. This is because these types of methods transform the raw features into binary features, which can effectively reduce the computation time compared with others. Moreover, SDMFC achieves a much better computational efficiency than the other hand-crafted methods of Compcode and GSLGS. We believe this to be the case since the projection functions of the proposed method are first learned and saved through the training samples, and then directly used to extract the features of the testing samples, thereby greatly reducing the time cost. In addition, the proposed SDMFC jointly learns the multimodal features and projects it into discriminant binary codes by feature mapping, which can significantly improve the computational efficiency. Therefore, the proposed SDMFC possesses a competitive computational speed as well as a better recognition accuracy compared with the other methods.

E. Parameters Sensitivity Analysis

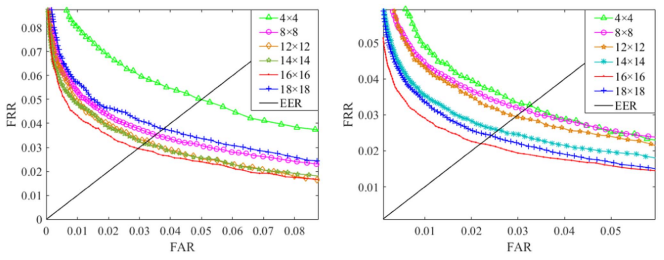
λ_1, λ_2 and λ_3 are three trade-off parameters (refer to (4) of the proposed SDMFC that leverage the importance of each constraint. To select the suitable parameters of SDMFC, we evaluated the effect of these variables on the three multimodal finger databases and compared the AVE results by setting different parameters. The analysis of the parameters is conducted by varying one variable value at a time and fixing the value of the other variables. Specifically, we firstly fixed $\lambda_2 = 10$, λ_3=0.01, and parameterized λ_1 by a discrete set [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000]. Then, we fixed λ_1=1, λ_3=0.01, and parameterized λ_2. Lastly, we fixed λ_1=1, λ_2=10, and parameterized λ_3. Table IV reports the corresponding identification accuracy versus parameter λ_1 on the three multimodal finger databases. It can be seen from Table IV that SDMFC conformably achieves a best identification performance when λ_1=1. In the Data-multi database, the identification accuracy rate varies slightly with the change of variable λ_1. In the SD-PolyU and USM-PolyU databases, the identification accuracy accelerates with the increasing values of λ_1 until λ_1=1, at which point there is no more increase. The main reason is due to the fact that the small values of λ_1 limit the contribution of the sparse constraint term in SDMFC. Consequently, the projection matrix has a low discriminative capacity. As the value of λ_1 increases, the sparse constraint term has more power, thereby bringing a better identification accuracy.

The parameters λ_2 and λ_3 of SDMFC are used to balance the two constraints of the intra-modality discriminant term and the inter-modality correlation term. Fig. 6 illustrates the effect of different λ_2 and λ_3 for SDMFC, respectively. By observing these identification results from Fig. 6(a), we can see that SDMFC achieves a stable identification accuracy when λ_2 is chosen between [0.01, 10]. The value of λ_2 that is too large or too small will degrade the identification performance of SDMFC.



(a) (b)

Fig. 6. The average identification accuracy versus parameters λ_2 and λ_3. (a) Different λ_2 values. (b) Different λ_3 values.



(a) (b)

Fig. 7. The ROCs curves of different division block sizes on the (a) SD-PolyU and (b) USM-PolyU.

This demonstrates that the intra-modality complementary learning can improve the performance of SDMFC. From Fig. 6(b), we can find that the recognition performance is relatively stable on the Data-multi database when λ_3 is located in the range of [0.0001, 10000]. As for the SD-PolyU and USM-PolyU databases, as λ_3 increases, the average accuracy of SDMFC slightly increases as well at first before a decrease in accuracy when $\lambda_3 > 0.01$. One possible reason is that the finger images in the Data-multi database have smaller within-class changes than the other databases, making the recognition accuracy insensitive to changes in the variables.

In the Section III-D, the learned feature maps are uniformly divided into non-overlapping sub-blocks. Hence, the size of division blocks also affects the recognition performance of the proposed method. To choose the appropriate division block sizes, we compared the EERs of different block sizes in the feature maps. Fig. 7 depicts the ROC curves of different division block sizes on the SD-PolyU and USM-PolyU databases, respectively. From Fig. 7, we can observe that the division block with a size of 16×16 obtains the lowest EER on the two multimodal finger databases.

F. Ablation Study

For the proposed SDMFC, we combined the sparse constraint term, intra-modality discriminative term, and inter-modality correlation term for multimodal feature learning. To comprehensively evaluate the discriminative ability of SDMFC, we performed an ablation study for the three constraints on the SD-PolyU and USM-PolyU databases. In this subsection, the experiment is carried out by setting the parameters λ_1, λ_2, and λ_3 to 0, respectively. Table V lists the average identification accuracy resulting from different constraints. It can be clearly seen

TABLE V
INFLUENCES OF THREE CONSTRAINTS ON THE SD-POLYU AND USM-POLYU DATABASES

Methods	SD-PolyU	USM-PolyU
SDMFC$_{\lambda_1=0}$	92.8720±2.2917	95.0364±0.2948
SDMFC$_{\lambda_2=0}$	92.6800±2.3673	95.8000±0.5676
SDMFC$_{\lambda_3=0}$	92.2160±2.8808	95.0727±0.2750
SDMFC	**97.4240±1.7629**	**97.7205±0.5469**

TABLE VI
THE IDENTIFICATION RESULTS (%) OF THE PROPOSED METHOD ON THE FIVE UNI-MODAL DATABASES

Databases	AVE(%)
SDMFC_Data-fv	99.2199±0.1052
SDMFC_Data-fkp	98.3514±0.2664
SDMFC_SDUMLA-fv	82.4080±2.7487
SDMFC_FV-USM	89.4045±1.1623
SDMFC_PolyU-fkp	87.2864±0.8176

from Table V that without any of these three constraints, the identification performance degrades. This result demonstrates that the three constraint terms of SDMFC contribute to improving the identification performance. Significantly, the relative contributions of different constraints are different. In detail, the accuracy of SDMFC without inter-modality constraint ($\lambda_3=0$) is lower than that of without intra-modality constraint ($\lambda_2=0$). This indicates that the inter-modality correlation term has more influence than the intra-modality discriminant term overall on the identification performance of SDMFC.

In addition, to verify the effectiveness of multimodal features in SDMFC, we conducted an ablation study on uni-modal and multimodal finger recognition. We set $\lambda_1=1$, $\lambda_2=10$, and $\lambda_3=0$ to learn the finger features of a single modality. Table VI lists the identification results on the five uni-modal finger databases. From Table VI, we can find that the identification accuracy of SDMFC still achieves 99.2199% on Data-fv and 98.3514% on Data-fkp, respectively. This verifies the superior feature representation ability of the proposed method. Moreover, by comparing uni-modal and multimodal recognition results, we can observe that the multimodal finger recognition always achieves a better accuracy than uni-modal finger recognition. The main reason is that multimodal recognition can fully take into account the discrimination and latent relationships among different modalities. This further demonstrates the feature representation capability of multimodal recognition in SDMFC.

V. CONCLUSION

In this paper, we explored a sparsity and discriminant feature codes learning approach and applied it for multimodal finger recognition. Different from most conventional multimodal biometrics technologies that focused on extracting information from multiple modalities equally and independently, our proposed SDMFC jointly learned the specific and common information among the inter-modality and intra-modality samples. More importantly, our proposed SDMFC adaptively learned a bank of collective mapping functions to project different modalities data into a latent common subspace. For the procedure of

SDMFC, we first established a local difference matrix to describe the rich texture information of the given finger images. Then, we projected the multimodal features into sparse and discriminative binary codes by the pre-trained mapping functions. Lastly, we developed a SDMFC-based block-wise histogram descriptor for feature representation. Extensive experimental results on three multimodal finger databases demonstrated that the proposed SDMFC achieved a state-of-the-art recognition performance for multimodal finger recognition. The results further show the effectiveness of the binary code learning-based methods in multimodal biometric recognition.

In future work, we will explore the application of SDMFC in other modalities, including face, iris, fingerprint, and palm vein, etc. Furthermore, we will explore a non-negative low-rank feature learning approach for multimodal biometric recognition.

REFERENCES

[1] M. Jian and K. Lam, "Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1761–1772, Nov. 2015.

[2] M. Jian, K. Lam, and J. Dong, "A novel face-hallucination scheme based on singular value decomposition," *Pattern Recognit.*, vol. 46, pp. 3091–3102, 2013.

[3] L. Fei *et al.*, "Learning compact multifeature codes for palmprint recognition from a single training image per palm," *IEEE Trans. Multimedia*, vol. 23, pp. 2930–2942, 2021, doi: 10.1109/TMM.2020.3019701.

[4] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1 4, no. 1, pp. 4–20, Jan. 2004.

[5] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 113–126, Jan. 2014.

[6] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, pp. 2049–2058, 2015.

[7] N. E. D. Elmadany, Y. He, and L. Guan, "Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 21, pp. 1317–1331, 2019.

[8] J. Yang, Y. Shi, and G. Jia, "Finger-vein image matching based on adaptive curve transformation," *Pattern Recognit.*, vol. 66, pp. 34–43, 2017.

[9] A. Kumar and Y. Zhou, "Human identification using finger images," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2228–2244, Apr. 2012.

[10] L. Zhang, L. Zhang, D. Zhang, and H. Zhu, "Online finger-knuckle-print verification for personal authentication," *Pattern Recognit.*, vol. 43, no. 7, pp. 2560–2571, 2010.

[11] L. Fei *et al.*, "Jointly learning compact multi-view hash codes for few-shot FKP recognition," *Pattern Recognit.*, vol. 115, 2021, Art. no. 107894.

[12] S. Li, B. Zhang, L. Fei, and S. Zhao, "Joint discriminative feature learning for multimodal finger recognition," *Pattern Recognit.*, vol. 111, Mar. 2021, Art. no. 107704.

[13] S. Li and B. Zhang, "Joint discriminative sparse coding for robust hand-based multimodal recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 3186–3198, 2021.

[14] J. Yang, X. Zhang, "Feature-level fusion of fingerprint and finger-vein for personal identification," *Pattern Recognit. Lett.*, vol. 33, pp. 623–628, Apr. 2012.

[15] W. Yang, Z. Chen, C. Qin, and Q. Liao, "α -trimmed Weber representation and cross section asymmetrical coding for human identification using finger images," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 1, pp. 90–101, Jan. 2019.

[16] J.-D. Wu and C.-T. Liu, "Finger-vein pattern identification using principal component analysis and the neural network technique," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5423–5427, 2011.

[17] J. Wen *et al.*, "Robust sparse linear discriminant analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 390–403, Feb. 2019.

[18] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[19] J. Yang, Y. Shi, and J. Yang, "Finger-vein recognition based on a bank of Gabor filters," in *Proc. 10th Asian Conf. Comput. Vis.*, Queenstown, New Zealand: ACCV, 2010, pp. 374–383.

[20] W. Yang, C. Hui, Z. Chen, J. Xue, and Q. Liao, "FV-GAN: Finger vein representation using generative adversarial networks," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 9, pp. 2512–2524, Sep. 2019.

[21] S. Li, B. Zhang, S. Zhao, and J. Feng, "Local coding based convolutional feature representation for multimodal finger recognition," *Inf. Sci.*, vol. 547, pp. 1170–1181, Feb. 2021.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, Sep. 2014 . [Online]. Available https://arxiv.org/abs/1409.1556

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 770–778.

[24] J. Shen and N. Robertson, "BBAS: Towards large scale effective ensemble adversarial attacks against deep neural network learning," *Inf. Sci.*, vol. 569, pp. 469–478, 2021.

[25] M. Jian and K. Lam, "Face-image retrieval based on singular values and potential-field representation," *Signal Process.*, vol. 100, pp. 9–15, 2014.

[26] M. Jian *et al.*, "Multi-view face hallucination using SVD and a mapping model," *Inf. Sci.*, vol. 488, pp. 181–189, 2019.

[27] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 24–38, Jan. 2016.

[28] A. Abdi, M. Rahmati, and M. M. Ebadzadeh, "Entropy based dictionary learning for image classification," *Pattern Recognit.*, vol. 110, 2021, Art. no. 107634.

[29] C. Yang, J. Shen, J. Peng, and J. Fan, "Image collection summarization via dictionary learning for sparse representation," *Pattern Recognit.*, vol. 46, pp. 948–961, 2013.

[30] I. Rida, S. A. Maadeed, A. Mahmood, A. Bouridane, and S. Bakshi, "Palmprint identification using an ensemble of sparse representations," *IEEE Access*, vol. 6, pp. 3241–3248, 2018.

[31] S. Zeng, B. Zhang, J. Gou, and Y. Xu, "Regularization on augmented data to diversify sparse representation for robust image classification," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2020.3025757.

[32] L. Xie, J. Shen, J. Han, L. Zhu, and L. Shao, "Dynamic multi-view hashing for online image retrieval." in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 3313–3319.

[33] D. Wang, X. Gao, X. Wang, and L. He, "Label consistent matrix factorization hashing for large-scale cross-modal similarity search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2466–2479, Oct. 2019.

[34] S. Li and B. Zhang, "An adaptive discriminant and sparsity feature descriptor for finger vein recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 2140–2144.

[35] H. Liu, R. Ji, Y. Wu, F. Huang, and B. Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7380–7388.

[36] L. Fei *et al.*, "Joint multi-view feature learning for hand-print recognition," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9743–9755, Dec. 2020.

[37] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1657–1663, Jun. 2010.

[38] S. Li, H. Zhang, Y. Shi, and J. Yang, "Novel local coding algorithm for multimodal finger feature description and recognition," *Sensors*, vol. 19, no. 9, 2019, Art. no. 2213

[39] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2041–2056, Oct. 2015.

[40] J. Lu, V. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1979–1993, Aug. 2018.

[41] L. Fei *et al.*, "Learning discriminant direction binary palmprint descriptor," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3808–3820, Aug. 2019.

[42] B. Hu *et al.* "A lightweight spatial and temporal multi-feature fusion network for defect detection," *IEEE Trans. Image Process.*, vol. 30, pp. 472–486, 2021.

[43] X. Zhang, X. Gao, W. Lu, L. He, and J. Li, "Beyond vision: A multimodal recurrent attention convolutional neural network for unified image aesthetic prediction tasks," *IEEE Trans. Multimedia*, vol. 23, pp. 611–623, 2021.

[44] L. Zhang *et al.* "Multimodal marketing intent analysis for effective targeted advertising," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2021.3073267.
[45] W. M. Yang, X. L. Huang, F. Zhou, and Q. M. Liao, "Comparative competitive coding for personal identification by using finger vein and finger dorsal texture fusion," *Inform. Sci.*, vol. 268, pp. 20–32, 2014.
[46] W. Yang, S. Wang, and J. Hu, " A fingerprint and finger-vein based cancelable multi-biometric system," *Pattern Recognit.*, vol. 78, pp. 242–251, Jun. 2018.
[47] H. Zhang, S. Li, Y. Shi, and J. Yang. "Graph fusion for finger multimodal biometrics," *IEEE Access*, vol. 7, pp. 28607–28615, 2019.
[48] Y. Luo *et al.* "Local line directional pattern for palmprint recognition," *Pattern Recognit.*, vol. 50, pp. 26–44, 2016.
[49] Y. L. Yin, L. L. Liu, and X. W. Sun, "SDUMLA-HMT: A multimodal biometric database," in *Proc. Chin. Conf. Biometric Recognit.*, 2011, pp. 260–268.
[50] M. S. M. Asaari, S. A. Suandi, and B. A. Rosdi, "Fusion of band limited phase only correlation and width centroid contour distance for finger based biometrics," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3367–3382, 2014.

Shuyi Li received the M.S. degree from the Civil Aviation University of China, Tianjin, China, in 2016. She is currently working toward the Ph.D. degree in computer science from the PAMI Research Group, Department of Computer and Information Science, University of Macau, Macau, China. Her research interests include biometrics, pattern recognition, image processing, and machine learning.

Bob Zhang (Senior Member, IEEE) received the B.A. degree in computer science from York University, Toronto, ON, Canada, in 2006, the M.A.Sc. degree in information systems security from Concordia University, Montreal, QC, Canada, in 2007, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2011. After graduating from Waterloo, he remained with the Center for Pattern Recognition and Machine Intelligence, and later was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently an Associate Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include biometrics, pattern recognition, and image processing. Dr. Zhang is a Technical Committee Member of the IEEE Systems, Man, and Cybernetics Society and an Associate Editor for the *IET Computer Vision*.

Lunke Fei (Member, IEEE) received the Ph.D. degree in computer science and technology from the Harbin Institute of Technology, Harbin, China, in 2016. Since April 2017, he has been with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. His research interests include pattern recognition, biometrics, image processing, and machine learning.

Shuping Zhao (Member, IEEE) received the M.S. degree in information science from South China Normal University, Guangzhou, China, and Ph.D. degree in computer science from the University of Macau, Macau, China, in 2021. Since 2021, he has been with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. His research interests include deep learning and linear representation for palmprint recognition.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA. He is currently an Associate Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security. He is a Senior Member of the International Society for Optical Engineering (SPIE). He was the recipient of the Third Price of Macau Natural Science Award in 2014. He is the Co-Chair of Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He is an Associate Editor for the IEEE TRANSACTIONS ON NEUTRAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and four other journals.